QUALITY ASSURANCE FOR HIGH-THROUGHPUT BIOASSAY METHODS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority under 35 U.S.C. § 119(e) to U.S. Provisional Patent Application No. 60/389,831, entitled "Quality Assurance/Quality Control for SELDI-TOF Mass Spectra," filed on July 29, 2002, the contents of which are hereby incorporated by reference in their entirety.

STATEMENT OF FEDERALLY SPONSORED RESEARCH

[0002] The research performed in connection with some of the subject matter disclosed in this application was performed on samples supplied by the United States Government.

BACKGROUND OF THE INVENTION

[0003] The invention relates generally to the field of bioinformatics. More specifically, the invention relates to a method of quality control for bioinformatic systems.

[0004] Methods of analyzing biological samples are generally known. In a typical analysis, mass spectroscopy is performed on the biological sample to determine its overall biochemical make-up. Based on the mass spectra obtained from the mass spectroscopy, various diagnostics may be run.

[0005] When biological samples are analyzed, it is desirable to perform more than one trial on the biological sample, thereby improving the accuracy of the diagnostic. Analysis of biological samples may be performed by using biochips, electrospray, or other protein separation techniques. A problem arises as samples are analyzed over time. The apparatus used to extract data from the samples may become uncalibrated, or a variation between chips when using biochips (e.g., in using SELDI or MALDI methods), or between diluents (when using electrospray techniques) may cause data obtained from the sample to become skewed.

[0006] Therefore, there is a need for a method of monitoring bioassay processes to determine when the process may be producing inaccurate data that may lead to a misdiagnosis.

SUMMARY OF THE INVENTION

[0007] The invention provides a quality control method for ensuring that a particular bioassay process is yielding acceptable data. Using the methods of the invention, the reliability of data used in a diagnostic procedure may be improved. Another embodiment of the invention includes using the Knowledge Discovery Engine ("KDE") to classify and archive biochips and to distinguish between type of biochips. Alternatively, the KDE may be used to classify and archive diluents and distinguish between diluents having different composition or concentrations.

[0008] The invention may use the KDE to identify hidden patterns across a wide variety of serum samples and biochips to generate a control model. Alternatively, the KDE does not have to be used to perform the methods of the invention.

[0009] The KDE is disclosed in U.S. Patent Application Serial No. 09/883,196, now U.S. Application Publication No. 2002/0046198A1, entitled "Heuristic Methods of Classification," filed June 19, 2001 ("Heuristic Methods"), and U.S. Patent Application Serial No. 09/906,661, now U.S. Application Publication No. 2003/0004402, entitled "A Process for Discriminating Between Biological States Based on Hidden Patterns from Biological Data," filed July 18, 2001 ("Hidden Patterns"), the contents of both of which are hereby incorporated by reference in their entirety. Software running the KDE is available from Correlogic Systems, Inc. under the name Proteome Quest™.

[0010] As described above, the KDE does not need to be used to practice the invention. One method of practicing the invention includes defining a number of features characteristic of the control sample. As used herein the term "feature" refers to a particular mass to charge ratio (m/z) within a spectrum. Additionally, as used herein, the term "vector" refers to a feature having a particular magnitude. Therefore, a vector is a two-dimensional value having both a mass to charge value and a magnitude.

[0011] After the features are defined, the vectors are plotted in n-dimensional space, where n is the number of defined features. The plotted vectors will define a centroid. A centroid is a

reference point located in n-dimensional space and associated with the selected features. This centroid is the control centroid.

[0012] Using a preserved aliquot of the same mixture used to generate the model, the spectral information from the aliquot of molecular mixture is obtained using the bioassay process. The vectors associated with the predetermined features selected in the generation of the control model are mapped in n-dimensional space. Using these newly mapped vectors, a comparison can be made of the deviation of a test centroid based on the newly mapped vectors from the control centroid associated with the control model.

[0013] When a determination is made that the test centroid deviates too far from the control centroid, calibration of the apparatus, a change of equipment, or other system adjustment may be needed. For example, if a general trend away from the control centroid is noticed, this will indicate that the mass spectrometer or other analyzer is uncalibrated or may need repair. A sudden shift in the data away from the control centroid may mean a number of things. First, it may mean that diluents or other preparation reagents were bad. In this case, a new diluent or reagent should be mixed. Additionally, a sudden shift could indicate that the type of biochip being used as a protein separation method has been changed. Alternatively, a sudden shift in the data away from the control centroid may indicate that the apparatus settings have changed, and the apparatus should be recalibrated.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] FIG. 1 is a flow chart of a method of obtaining a control model according to one aspect of the invention.

[0015] FIG. 2 shows a method of testing a system against the control model to determine an error in the control model.

[0016] FIG. 3A and 3B are exemplary spectra obtained using the methods of the invention.

[0017] FIG. 4 illustrates an example of comparing a test centroid to a control centroid according to one embodiment of the invention.

[0018] FIG. 5 depicts a method of generating a model for distinguishing between types of biochips according to an aspect to the invention.

[0019] FIG. 6 illustrates a method of comparing a new batch of biochips to an archive of biochips to classify types of biochips.

[0020] FIG. 7. is a trend plot illustrating the performance of an electrospray system when compared with a predetermined model.

[0021] FIG. 8 is a trend plot illustrating the performance of a SELDI system utilizing biochips when compared to a biochip control model.

DETAILED DESCRIPTION

[0020] Generally, the invention includes a method of obtaining a control model for use in a bioinformatics system and a method for quality assurance in a bioassay process. Another embodiment of the invention may be used to distinguish between different types of biochips or diluents used in electrospray processes.

[0021] A method of obtaining a control model according to an aspect of the invention is illustrated generally in FIG. 1. A mixture of molecules is prepared in a step 110. The mixture of molecules can include any mixture of molecules. The mixture of molecules may include any natural or artificial molecules. The molecules may have an atomic weight of greater than 400 and be water soluble. In one embodiment, the mixture of molecules can include a mixture of isolated peptides.

[0022] After the mixture of molecules is prepared, the mixture is divided into aliquots at a step 120. Enough aliquots may be made so as to permit continued use over a desired number of tests, which may be over a period of years. The aliquoted mixture may be used for comparison purposes after generation of the control model.

[0023] Once the mixture is divided into aliquots, all of the aliquots are preserved in step 130. In one embodiment, this includes freezing the mixture in liquid nitrogen. To enhance the consistency of the results, it is desirable to freeze all of the aliquots of the mixture so that any

change that the mixture may undergo due to the freezing or thawing process is constant across all aliquots.

[0024] Next, to obtain data using a bioassay process, some aliquots of the mixture are thawed. The aliquots may be thawed at room temperature. Once the mixture has thawed, the mixture may be placed in a ice bath at about 4°C. Although it is not critical that the mixture be kept at 4°C, it is advantageous because most biological samples have a high degree of stability around this temperature.

[0025] In one embodiment of the invention, the molecular mixture is then analyzed using mass spectrometry, and a data set based on the mass spectrometry is obtained at a step 140.

[0026] In an embodiment where the KDE is not used, a selection of features may be made from the mass spectrometry data at step 150. For example, every thousandth feature may be selected, as illustrated in FIG. 3A. Once the features are selected, these features are fixed, i.e., these will be the features that will be observed for all test samples. Any number of features may be selected, for example, every tenth feature, every hundredth feature, or every thousandth feature. The features chosen may be completely random as well, as long as the same features are observed and compared against the control model during testing.

[0027] Once the features are chosen, vectors based on these features are mapped into n-dimensional space, where "n" corresponds to the number of features selected, to define a centroid in that space at step 160. The vectors have mass to charge values representing the selected features, and have a magnitude. The location of the centroid may define the control model. This centroid may act as the basis for comparison for the test spectra. [The centroid is a point identifying the center of all the vectors associated with the chosen features in n-dimensional space.]

[0028] Any number of aliquots of the mixture may be analyzed to generate the control model. Preferably, more than one aliquot of the mixture is analyzed. The greater the number of aliquots analyzed, the more robust the model will be. When analyzing multiple aliquots, the features observed should always be the same features (e.g., every thousandth feature). This is illustrated

in FIG. 3B. FIGS. 3A and 3B are examples of mass spectra outputs from a hypothetical molecular mixture. One mass spectrum that was used to generate the control model is shown in FIG. 3A. Every thousandth feature has been selected, yielding a total of 20 features ($F_1 - F_{20}$).

[0029] An exemplary test sample is shown in FIG. 3B, having the same features selected (i.e., every thousandth feature is selected). While substantially the same, the spectrum illustrated in FIG. 3A is different from that illustrated in FIG. 3B in that some of the peaks have a different magnitude. For example, peak F_{10} has a slightly greater magnitude in FIG. 3B. Peaks F_1, F_{14}, and F_{20} have smaller magnitudes in FIG. 3B than those shown in FIG. 3A. These varied magnitudes will impact the location of the vector associated with that feature in n-dimensional space. By producing features with different magnitudes (i.e., a vector), the centroid based on those features will be displaced with respect to the control centroid.

[0030] A method of using the control model for quality assurance/quality control according to an aspect of the invention is illustrated generally in FIG. 2. Once the control model is generated, spectra may be compared against the control model to determine if the system, the biochip, or the diluents being used are yielding precise results.

[0031] To run a diagnostic on the system, an aliquot of the initial molecular mixture may be retrieved and thawed in a step 210. The thawing process should replicate the thawing process used for generating the control model. For example, if the aliquots used for making the control model were thawed at room temperature, then the aliquot used to run the test model should be thawed at room temperature. If the aliquots used to generate the control model were then placed in an ice bath at 4°C, then the aliquot for the test model may be placed in an ice bath at 4°C. The consistency in retrieval methods may enhance precision and prevent errors that may be generated by different preservation and retrieval methods.

[0032] Then, using the same bioassay process used to obtain the control model, spectral data may be obtained for the mixture in a step 220. To achieve a basis for comparison, the same features should be identified in the test spectrum. Vectors associated with these features may be mapped in n-dimensional space. The set of vectors obtained from this mapping will define a test

centroid in step 240. The test centroid may then be used to determine the degree of error between the test spectrum and the control spectrum in a step 250.

[0033] Error may be determined by calculating a distance between the fixed control centroid and the test centroid. The distance may be calculated in n-dimensional space as described in further detail below. An acceptable degree of error may be one to two standard deviations. The standard deviation is calculated based on the vectors used to develop the control centroid. If the test centroid is greater than a predetermined distance from the control centroid, a problem may exist in the apparatus used to conduct the bioassay process (e.g., the biochip surfaces, the electrospray apparatus) or the diluents used (with respect to the electrospray processes).

[0034] A method of determining the error according to an aspect of the invention is shown generally in FIG. 4. While the plot shown in FIG. 4 is depicted in three dimensions, it should be understood that a plot obtained by using the methods of the invention could be in any number of dimensions. A three-dimensional model is illustrated because it is easily conceptualized.

[0035] FIG. 4 depicts a first sphere in three-dimensional space, F_M. Sphere F_M contains a set of all vectors corresponding to the selected features. This concept could be conceptualized as a sphere containing (on its surface and/or in its interior) a multitude of points. The plotted vectors need not be homogeneously distributed thought the sphere (or in n-dimensional space the hypersphere or other hyper-volume). The plotted vectors may be averaged to a centroid, "M." The centroid will be located at the center of the sphere. The centroid "M" is the control model centroid, and will serve as the basis of comparison for test centroids.

[0036] When a test spectrum is obtained, the test spectrum may also be plotted. The vectors for a single spectrum would plot to a single point in the illustrated space. Vectors for multiple spectra would plot to a point for each spectrum. The plotting of multiple spectra is again, for ease of conceptualization, illustrated in FIG. 4 as a sphere, F_T, but one of skill in the art will understand that the number of dimensions will be dependent on the number of features selected in generating the control model. Sphere F_T is centered on a test centroid, designated as "T." As can be seen in FIG. 4, the spheres defining the features are offset, in that their centroids are not colocated at the same point in the space. If the spheres were centered on the same point, then

one would know that the test model was identical with the control model. In some instances, this may not be the case. Many times, the test centroid, "T" will be displaced from the control model centroid, "M." This displacement is an indicator of the status of the bioassay process as a whole. The displacement between the two centroids is illustrated generally as "d." In three dimensional space, "d" may be determined using the following three-dimensional formula for the distance between the two centroids in FIG. 4 is:

$$d = |M(x,y,z)T(a,b,c)| = \sqrt{(x-a)^2 + (y-b)^2 + (z-c)^2} .$$

[0037] When expanding this to n-dimensional space, the distance between the two points is given by:

$$d = |M(m_1, m_2, ..., m_n)T(t_1, t_2, ..., t_n)| = \sqrt{\sum_{i=1}^{n} (m_i - q_i)^2} .$$

[0038] If the distance between points "M" and "T" are greater than a predetermined tolerance, then calibration or other system changes may be needed (e.g., new diluents may need to be used, new biochips etc.). If the distance between points "M" and "T" are equal to or within the predetermined tolerances, then samples may be evaluated using the high-throughput bioassay process.

[0039] Applying the method of the invention to an electrospray system requires only a few modifications to the method as applied to biochip technology. For example, in generating the control model, as illustrated in FIG. 1, step 140 would include obtaining data using an electrospray process for protein separation (rather than using biochips). The remaining steps are substantively unchanged.

[0040] The invention may be used to monitor the performance of an overall system, and depending on the behavior of the test model with respect to the control model, one may determine whether a particular aspect of the system is producing unreliable results.

[0041] One method of monitoring system performance is to generate a plot of the distance of the test centroid to the control centroid, using data taken over time. Exemplary plots are illustrated

in FIGS. 7 and 8. Monitoring these plots will allow overall system performance to be monitored. Furthermore, interpreting these plots will enable one to understand where problems may be arising that affect the quality of the system.

[0042] For example, if a general trend away from the control centroid is noticed, this will indicate that the mass spectrometer or other analyzer is uncalibrated, and may need repair. A sudden shift in the data away from the control centroid may mean a number of things. First, it may mean that diluents or other preparation reagents were bad. In this case, a new reagent should be mixed. Alternatively, a sudden shift in the data away from the control centroid may indicate that the apparatus settings have changed, and the apparatus should be recalibrated.

[0043] FIG. 7 illustrates a trend plot for monitoring an electrospray system. This plot illustrates a consistent system behavior. If anything, this plot illustrates that the system is settling in over time, thereby producing more consistent results. Here, the data is falling within one standard deviation of the mean, the mean being based on the mean distance of a vector in the control model from the control centroid.

[0044] FIG. 8 illustrates a trend plot for monitoring a SELDI system. The jump in the distance to centroid illustrated in FIG. 8 indicates a position where a new type of biochip was being used in the system. While this biochip performed within two standard deviations of the mean, it performed comparatively worse than the first biochip when being compared to the control model. The overall jump in the trend plot is due to the different chip surfaces which will produce different vectors in the test model. Because the different chip surface produces different vectors in the test model, the test centroid begins to drift away from the control model. In one embodiment of the invention, if the test centroid falls outside of two standard deviations from the mean, the system is deemed unsuitable. In another embodiment, the tolerances could be more strict, and the system may be deemed unsuitable if the test centroid was more than one standard deviation from the mean.

[0045] The methods of the invention may, in a particular application, be employed to determine if biochip surfaces are of a particular type and determined if the biochips are acceptable for use

in medical diagnostics. Furthermore, the KDE may be used to discover features that are only salient to different chip surfaces or different diluents used for electrospray systems.

[0046] A method of generating a model for classifying biochip type is shown generally in FIG. 5. As an initial step, a mixture of molecules is generated at a step 510. The mixture of molecules may be similar to that used to generate the model described above. The mixture of molecules is divided into aliquots at a step 520. The aliquots may then be preserved and retrieved at a step 530 in a manner similar to the method described in reference to FIG. 1. The mixture is then analyzed at a step 540 to obtain mass spectral data. The data obtained are then archived with respect to chip type. The data obtained from the group of biochips may then be input into the KDE at a step 550 to obtain a general biochip model. The KDE may extract only those features that are salient to distinguishing one biochip surface from another.

[0047] In general, the KDE will search for hidden or subtle patterns of molecular expression that are, in and of themselves, "diagnostic." The level of the identified molecular products is termed *per se* diagnostic, because the level of the product is diagnostic without any further consideration of the level of any other molecular products in the sample. In some instances, a normalizing molecular product may be used to normalize the level of the molecular products. The data may be normalized internally within the features in a vector. Alternatively, a synthetic peptide or other high molecular weight molecule may be added as an internal standard.

[0048] In the data cluster analysis utilizing the KDE, the diagnostic significance of the level of any particular marker, *e.g.*, a protein or transcript, is a function of the levels of the other elements that are used to calculate a sample vector. Such products are referred to as "contextual diagnostic products." The KDE's learning algorithm discovers wholly new classification patterns without knowing any prior information about the identity or relationships of the data pattern, *i.e.*, without prior input that a specified diagnostic molecular product is indicative of a particular classification.

[0049] As used in one method of the invention, data from each of the mass spectra are input into the KDE. The KDE then seeks to identify clusters of data (hidden patterns) in n-dimensional space, where n is the number of mass to charge values from the spectra, and each spectrum can

be mapped into the n-dimensional space using the magnitude of each of the selected mass to charge values in the spectrum. The KDE seeks clusters that are contain as many of the vectors as possible and that distinguish each of the biochips from the others.

[0050] After the model is generated, new biochips may be obtained, for example, from a newly manufactured batch of biochips. A method of classifying and identifying particular types of biochips is illustrated generally as FIG. 6. Once new biochips are obtained in step 600, it may be desirable to determine if the biochips satisfy predetermined standards for quality.

[0051] A number of biochips may be selected from the new group of biochips, and data obtained on the selected biochips at step a 610. To obtain data on the new biochip, an aliquot of the preserved mixture of molecules is retrieved. To ensure an accurate characterization, the molecular mixture to be used is as similar as possible, and preferably identical, to the original mixture used to construct the models. The data may be obtained using mass spectrometry.

[0052] Once data have been obtained on the selected biochip, the data are mapped in n-dimensional space to the model in a step 620. After the data is mapped to the model, a determination is made in a step 630 of whether the data maps to an archived biochip. If the data obtained from the mass spectrometry map to an archived biochip, the biochip being tested can be classified as the archived biochip in a step 631.

[0053] If the data do not map to an archived biochip, a determination is made in a step 640 of whether the data map consistently to a new cluster. If the data do map to a new cluster, the conclusion is reached in a step 641 that the biochip is of a new biochip type, and the biochip may be archived in a step 642.

[0054] If the data do not map to a new cluster, the data may map to a number of unrelated clusters in a step 650. If this is the case, the conclusion is reached in a step 660 that the batch of new biochips may be substandard, and a new batch should be obtained.

[0055] While the foregoing example was discussed with respect to a method of quality assurance/quality control for a SELDI- or other MALDI-type system, the methods of the invention are equally applicable to electrospray systems as well. For example, rather than

distinguishing between chip surfaces, the method when employed using electrospray is capable of distinguishing between different diluents or diluent concentrations.

[0056] A method for distinguishing between different types of diluent according to one aspect of the invention will now be described. Preparation, preservation, and retrieval of the mixture of molecules will be substantially the same as that described above with respect to biochips.

[0057] One particular difference is in the means of obtaining data from the molecular mixture. Using the electrospray system in conjunction with a mass spectrometer, mass spectral data is obtained. In order to obtain data using an electrospray, the mixture of molecules is combined with a diluent to achieve a predetermined concentration. Thus, the resultant spectral information will include information specific to the mixture of molecules, as well as spectral information specific to the diluent used. Therefore, rather than obtaining spectral data from multiple biochips and archiving the biochips, as illustrated in FIG. 5, step 540, one may obtain spectral data from multiple diluents and archive the diluents.

[0058] Another difference between the method of distinguishing between biochips and diluents includes obtaining new diluents, rather than obtaining new biochips, as illustrated in FIG. 6, 600. Additionally, rather than selecting random biochips, as shown in FIG. 6, 610, multiple samples of diluent may be used.

[0059] The method illustrated in FIG. 6 may be employed as illustrated. Using the method illustrated in FIG. 6 as applied to electrospray technology, diluents may be monitored for quality. If, for example, the data from the new diluent maps to a number of unrelated clusters, then the diluent may be substandard and a new diluent should be prepared.

[0049] While molecules disclosed had a molecular weight of greater than 400, other molecules having molecular weights acceptable for use in bioassay processes will be apparent to those skilled in the art based on the teachings provided herein.

[0050] As described above, preserving the aliquots was performed by freezing the aliquots using liquid nitrogen. The use of other cryogenic and non-cryogenic preservation methods are intended to be within the scope of the invention.

[0051] Additionally, one method of determining the distance between a test centroid and a control centroid is illustrated above for Cartesian systems. Other methods for calculating the distance between a control centroid and a test centroid are known in the art and are within the scope of the present invention. Some specific methods of calculating distance include Euclidian distance calculations, Hamming distance calculations, and Mahalanobis distance calculations

[0051] The various features of the invention have been described in relation to a method of quality assurance/quality control of high-throughput bioassay processes. However, it will be appreciated that many of the steps may be implemented with various apparatus and bioinformatics methods. Moreover, variations and modifications exist that would not depart from the scope of the invention.